JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

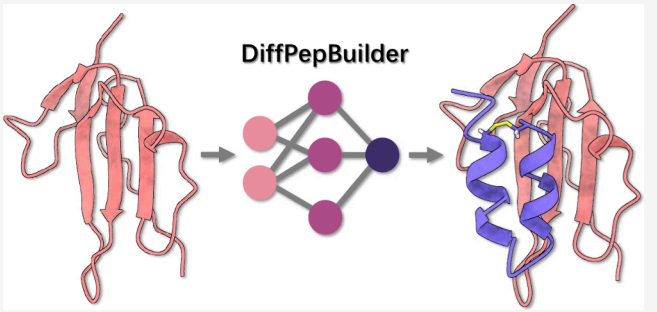
Target-Specific *De Novo* Peptide Binder Design with DiffPepBuilder

Fanhao Wang,[†] Yuzhe Wang,[†] Laiyi Feng, Changsheng Zhang,* and Luhua Lai*

Cite This: https://doi.org/10.1021/acs.jcim.4c00975 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Despite the exciting progress in target-specific *de novo* protein binder design, peptide binder design remains challenging due to the flexibility of peptide structures and the scarcity of protein-peptide complex structure data. In this study, we curated a large synthetic data set, referred to as PepPC-F, from the abundant protein−protein interface data and developed DiffPepBuilder, a *de novo* target-specific peptide binder generation method that utilizes an SE(3)-equivariant diffusion model trained on PepPC-F to codesign peptide sequences and structures. DiffPepBuilder also introduces disulfide bonds to stabilize the generated peptide structures. We tested DiffPepBuilder on 30 experimentally verified strong peptide binders with available protein−peptide complex structures. DiffPepBuilder was able to effectively recall the native structures and sequences of the peptide ligands and to generate novel peptide binders with improved binding free energy. We subsequently conducted *de novo* generation case studies on three targets. In both the regeneration test and case studies, DiffPepBuilder outperformed AfDesign and RFdiffusion coupled with ProteinMPNN, in terms of sequence and structure recall, interface quality, and structural diversity. Molecular dynamics simulations confirmed that the introduction of disulfide bonds enhanced the structural rigidity and binding performance of the generated peptides. As a general peptide binder *de novo* design tool, DiffPepBuilder can be used to design peptide binders for given protein targets with three-dimensional and binding site information.



1. INTRODUCTION

Peptides have emerged as promising candidates for drug development due to their diverse biological activities and relatively low toxicity.[1−5] Currently, over 100 peptide-based pharmaceuticals are used to treat various medical conditions, including cancer, diabetes, osteoporosis, multiple sclerosis, HIV, and chronic pain.[6−10] Semaglutide, a notable example recently, is a glucagon-like peptide-1 (GLP-1) receptor agonist used effectively in the management of type 2 diabetes and obesity, offering improved glycemic control and cardiovascular benefits.[11−13] The field of peptide drug design has also experienced significant progress, driven by advancements in computational methods,[14,15] structural biology,[16,17] and synthetic chemistry.[18−20] Computational tools and algorithms are pivotal in predicting peptide structures, interactions and pharmacokinetic properties, thus facilitating the development of peptide-based therapeutics. Traditional methods such as molecular dynamics simulations and docking studies have enabled researchers to explore the conformational space of peptides and predict their binding affinity to target proteins with improved accuracy.[21−25] For example, Chen et al.[26] has used Rosetta FlexPepDock[24] to design peptide scaffolds and sequences from existing peptides. However, traditional design methods encounter challenges in terms of efficiency and accuracy.

Peptide binder design remains challenging due to the flexibility of peptide structures and the scarcity of protein-peptide complex structure data. In this study, we curated a large synthetic data set, referred to as PepPC-F, from the abundant protein−protein interface data and developed DiffPepBuilder, a *de novo* target-specific peptide binder generation method that utilizes an SE(3)-equivariant diffusion model trained on PepPC-F to codesign peptide sequences and structures. DiffPepBuilder also introduces disulfide bonds to stabilize the generated peptide structures. We tested DiffPepBuilder on 30 experimentally verified strong peptide binders with available protein-peptide complex structures. DiffPepBuilder was able to effectively recall the native structures and sequences of the peptide ligands and generate novel peptide binders with improved binding free energy. We subsequently conducted *de novo* generation case studies on three targets. In both the regeneration test and case studies, DiffPepBuilder outperformed AfDesign and RFdiffusion coupled with ProteinMPNN, in terms of sequence and

Special Issue: Machine Learning in Bio-cheminformatics

Received: June 6, 2024
Revised: September 3, 2024
Accepted: September 4, 2024

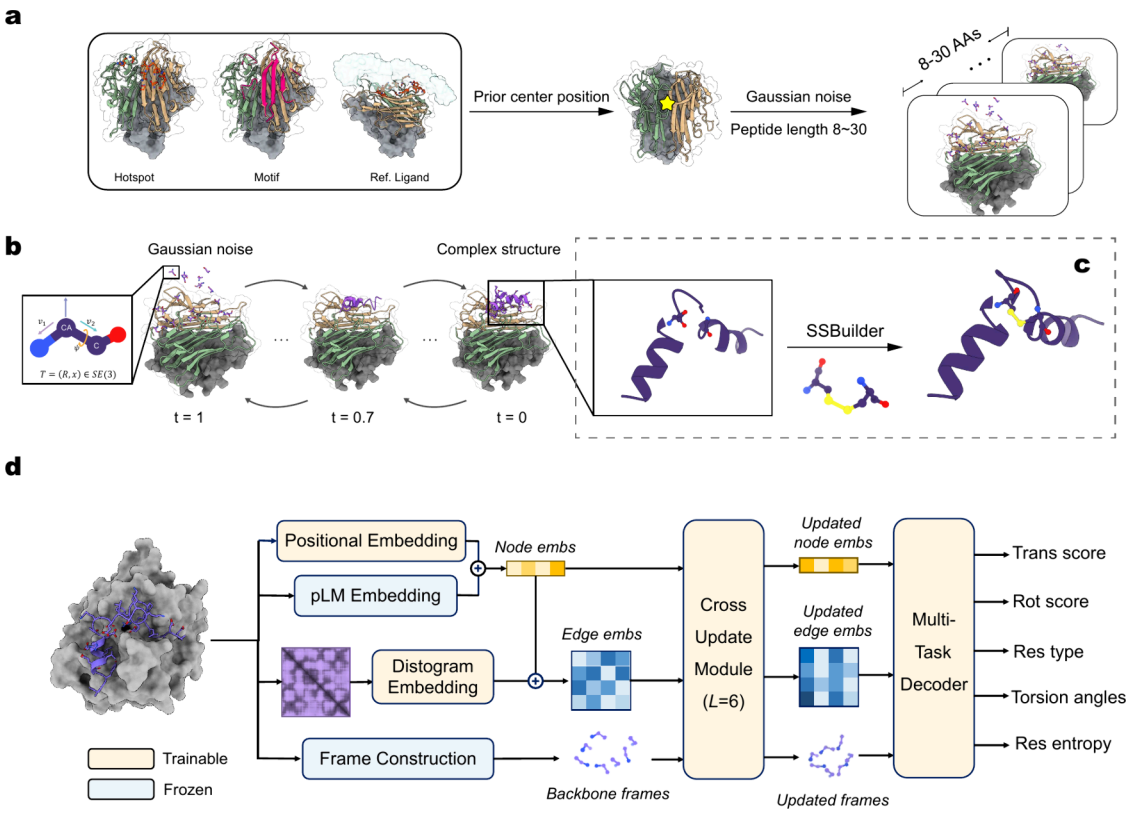


Figure 1. Overview of DiffPepBuilder. (a) Input preparation. DiffPepBuilder accepts user-specified binding site information formatted as Hotspots (in crimson), Motif (in crimson), or Reference ligand (transparent, used to define the hotspots). The model then determines the center position (marked by a yellow star) and length range (default 8 to 30) for generated peptide binders and initializes them with Gaussian noise. (b) The diffusion process. All residues in the model are represented by local frames that possess SE(3) equivariance. The coordinates of the target protein remain unchanged. The reverse (generative) process involves the peptide starting from Gaussian noise and progressively denoising into a complete binder structure. Conversely, the forward (noising) process entails the addition of Gaussian noise for model training. (c) Construction of disulfide bonds with the SSBuilder module. After generation, the peptide ligand structure, along with the residue entropy, is fed into SSBuilder. SSBuilder then filters out the residues whose entropy exceeds a threshold and constructs disulfide bonds among them based on geometrically matched fragments. (d) The architecture of the diffusion model. It primarily utilizes pLM Embeddings and positional encoding as node information, employs a distogram to additionally encode edge information, and converts three-dimensional coordinates into local frames. After these different modalities of information interact via the Cross Update Module, the Multitask Decoder outputs translational and rotational scores, predicted residue types, predicted torsion angles, and residue entropies.

structure recall, interface quality, and structural diversity. Molecular dynamics simulations confirmed that the introduction of disulfide bonds enhanced the structural rigidity and the binding performance of the generated peptides. As a general peptide binder *de novo* design tool, DiffPepBuilder can be used to design peptide binders for given protein targets with three-dimensional and binding site information.

Over the years, the Protein Data Bank (PDB)[27] has compiled a substantial amount of three-dimensional structural information for biomacromolecules and their complexes. Leveraging the expanding volume of available data, deep learning methods have achieved significant success in protein structure prediction[28−32] and design.[33−36] However, the number of well-defined protein-peptide complex structures in the PDB remains relatively limited. Peptides demonstrate a higher level of conformational plasticity with less secondary structures compared to proteins.[37−41] The flexibility of peptides and the scarcity of data pose challenges for deep learning models. Moreover, codesign of peptide binder backbone structures and sequences is particularly difficult due to the nonconservative nature of peptide backbones. Recently, geometric deep generative models, especially diffusion models[42−44] were used to meet these challenges. For instance, Kosugi et al.[45] developed the AfDesign[33] approach for the generation of backbone conformation and sequence features of peptide binders with good solubility. RFdiffusion[35] has demonstrated its capability in generating protein binders and also suggested potential adaptability for the generation of peptide ligands. Despite these advances, the application of deep generative models in peptide design highlights a pressing yet challenging opportunity, primarily confronted by three substantial hurdles: (1) the need for a comprehensive collection of protein-peptide complex structures for effective model training; (2) the difficult task of implementing codesign of structure and sequence; (3) the requirement of ensuring the stability of binding conformations of the generated peptides.

In this study, we first constructed a comprehensive protein-peptide structure data set from protein−protein interfaces and then developed a diffusion-based generative model to codesign backbone structures and sequences for peptide binders. The SE(3)-equivariant diffusion model architecture was further integrated with a geometric disulfide bond construction module to forge a novel tool referred to as DiffPepBuilder, which excels in peptide ligand sequence-structure codesign tasks. We tested DiffPepBuilder on 30 nonredundant protein-

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

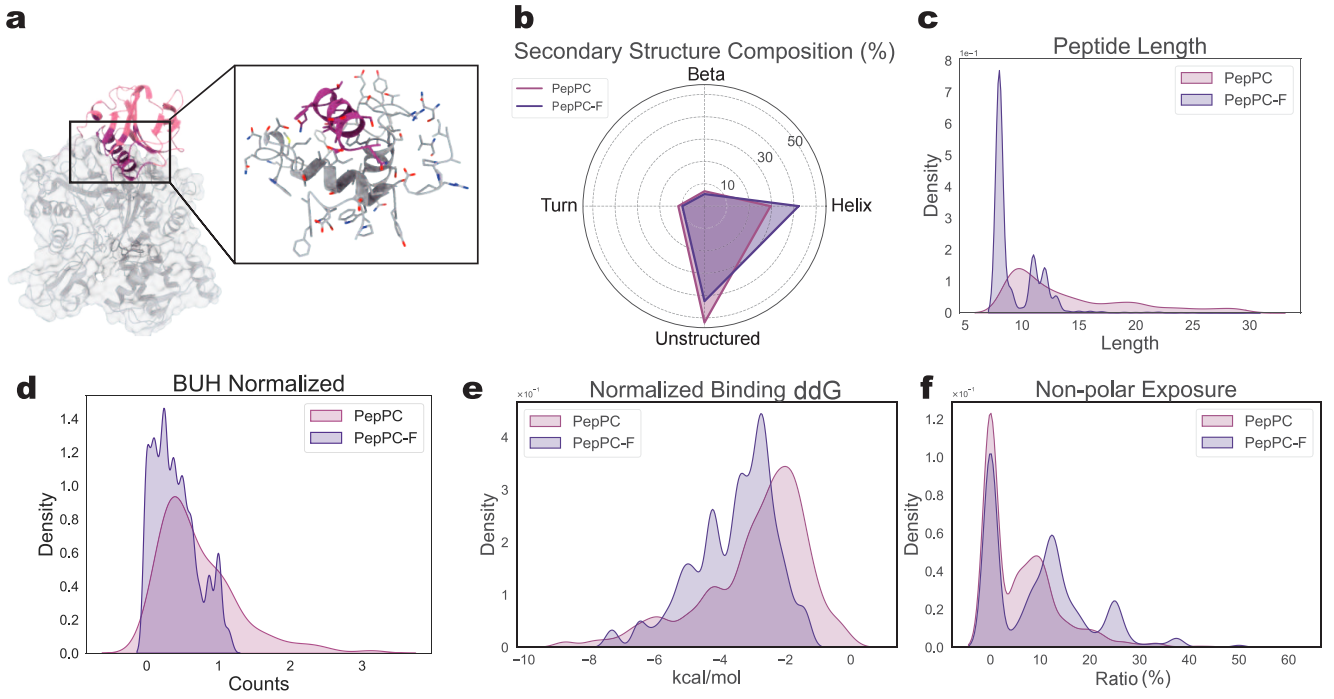


Figure 2. Statistics of the curated data sets. (a) An illustration of the construction process of protein-fragment complex entries is provided. Interface peptide fragments (in purple) are truncated from the complete binder (in pink) and, together with their corresponding binding proteins (in gray), form the complex structures that are collected in the PepPC-F data set. (b) A comparative diagram of secondary structures across two data sets, PepPC and PepPC-F. (c) A graph of the length distribution of PepPC and PepPC-F. (d) The number of unsatisfied hydrogen bonds at the interface normalized by residue count. (e) Distribution graph of Rosetta ddG normalized by residue count. (f) The ratio of ligand hydrophobic exposure residues between PepPC-F and PepPC.

peptide complex systems with strong binding strength and found that it can regenerate natural peptide binders with similar binding positions and high sequence similarity. We then used DiffPepBuilder to design novel linear and cyclic peptide binders for three important targets: activin-receptor-like kinase 1 (ALK1), 3C-like protease of SARS-CoV-2 ($3CL^{pro}$), and tumor necrosis factor (TNF-α). In both the regeneration tests and case studies, DiffPepBuilder outperformed AfDesign and RFdiffusion coupled with ProteinMPNN (RFd+MPNN).

2. RESULTS

2.1. Overview of the DiffPepBuilder Model. DiffPepBuilder is an end-to-end *de novo* peptide binder generation model that utilizes a diffusion-based generative procedure to simultaneously design the peptide sequence and structure through gradual structural denoising, coupled with a postprocessing procedure designed to enhance the stability of the generated peptide structures by introducing disulfide bonds when appropriate. The *de novo* peptide binder generation process is as follows: DiffPepBuilder takes the full target as input and initializes the protein binding pocket-peptide ligand complex structure based on user-specified binding site information and generation settings (Figure 1a, see section 4.2 for details). This initial complex structure is then fed into the diffusion-based generative procedure, as illustrated in Figure 1b, where an arbitrary residue is parametrized as an orientation-preserving rigid body frame $T_i \in \mathrm{SE}(3)$ as in AlphaFold 2[28] (SE(3) denotes the special Euclidean group). The full-atom structure of the generated peptide ligand is reconstructed by using the final ($t = 0$) frame representations. DiffPepBuilder subsequently identifies peptide residues whose amino acid type entropy exceeds a specified threshold, searches the disulfide bond fragment library for matching geometric conformations, and replaces the matched residues with disulfide bond-connected cysteine residues (Figure 1c, see section 4.4 for details). The structures of the peptide, both before and after the construction of disulfide bonds, are incorporated into the sampling repertoire.

The architecture of the denoising network is outlined in Figure 1d (see section 4.3 for details). The initial node and edge embeddings, in conjunction with the backbone frames, are processed through a 6-layer Cross Update Module adapted from FrameDiff[46] that alternately updates the embeddings and frame representations. A multitask decoder is utilized to predict the translational and rotational scores of the current diffusion step, the residue types of the peptide ligand, the main chain torsion angles ϕ/ψ, and the side chain torsion angles χ_1–χ_4. We employed a composite loss scheme (see section 4.5) and trained the denoising network exclusively on the PepPC-F (Peptide Protein Complexes-Fragment) data set (see section 2.5 for details).

2.2. Protein-Peptide Complex Data Set Construction and Analysis. Currently available protein-peptide complex structures are limited in quantity, and according to experimental data compiled from PDBbind2020,[47] peptide binders typically exhibit relatively weak binding affinity, mostly at the micromolar level, which is approximately 3 orders of magnitude weaker than protein–protein interactions (see Figure S1b). In order to build a deep learning model to design strong peptide binders, we used the protein–protein complex structures to build a large synthetic protein-peptide structure data set, PepPC-F. We collected structures of protein–protein complexes from PDB and extracted buried interfacial peptide fragments. The basic assumption is that

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

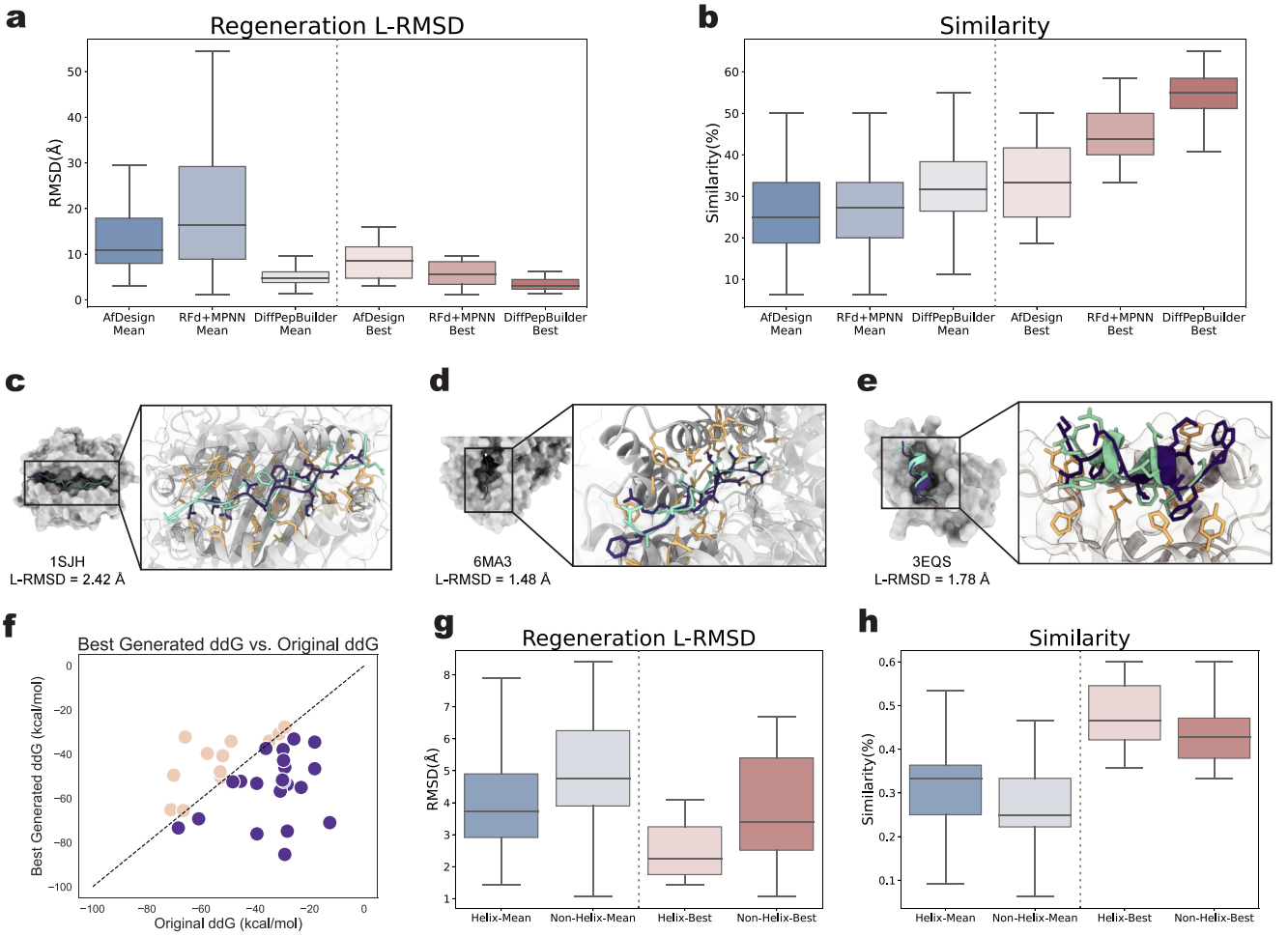


Figure 3. DiffPepBuilder's performance on regeneration test. (a) Distribution of mean and best L-RMSD values across different methods. (b) Distribution of mean and best similarity values across different methods. (c–e) Representative regeneration results for various targets. The targets (shown in gray, with interacting side chains in wheat) include human O-GlcNAc transferase, MHC class II, and MDM2. The original peptides are depicted in cyan, while the generated peptides are shown in violet. (f) Plot comparing the best ddG values of generated peptides to the ddG values of the original peptides in the testing targets. The best ddG refers to the lowest ddG achieved among all peptides generated for the same target. Targets for which the best generated ddG is lower than the original ddG are depicted in violet, while others are shown in wheat. (g and h) Distributions of L-RMSD and similarity values for peptides generated by DiffPepBuilder, categorized into helical and nonhelical groups.

peptide fragments that are exposed in the free ligand proteins and become buried by the target proteins represent good binding states of peptides. These peptide fragments, along with their corresponding binding proteins, constitute the complex structures collected in the PepPC-F data set. At the same time, we also built a data set of protein-peptide complexes, PepPC, for comparison. We restricted the peptide length in PepPC and PepPC-F data set between 8 and 30, akin to other databases such as ProPedia[48] and PepBDB.[49] We divided the peptides into helical (helix ratio ≥0.5) and nonhelical (including turns and unstructured peptides as shown in Figure 2b). After redundant removal and data cleaning (see section 4.1), there are 14,897 complexes in PepPC-F, including 4,241 helical peptides and 10656 nonhelical peptides and 3,832 complexes in PepPC, including 232 helical peptides and 3,600 nonhelical peptides. The size and diversity of PepPC-F make the training of the diffusion model feasible and enhance the models' generalization capability on various targets. We compared the characteristics of the peptides in these two data sets. The contents of secondary structures are similar to slightly high helical structures in PepPC-F (Figure 2b). Though both data sets contain peptides with 8–30 residues, short peptides are more populated in PepPC-F (Figure 2c). Peptides in PepPC-F form better interactions with the targeting proteins, as indicated by the fewer unsatisfied interface hydrogen bonds (Figure 2d,e) and better normalized binding free energy. One might worry that peptide fragments isolated from proteins may have a high proportion of exposed hydrophobic residues since protein interiors often feature tight hydrophobic packing, but our analysis shows that PepPC-F's hydrophobic exposure is not apparently different from that in the natural data set PepPC (Figure 2f). We also analyzed the proportion of hydrophilic and hydrophobic residues in both peptide data sets and observed a similar overall distribution. The PepPC-F data set is slightly more hydrophobic, but the difference from the natural data set is not significant (see Figure S1a).

2.3. DiffPepBuilder Regenerates Peptide Binders in Known Protein-Peptide Complexes. After training DiffPepBuilder on the PepPC-F data set, we first tested whether it can regenerate peptide binders in known protein-peptide complexes. For this purpose, we constructed an independent testing data set, PepPC-HF (PepPC High binding Affinity), that contains 30 nonredundant protein-peptide complexes with high-resolution structures (better than 2.5 Å) and strong

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

binding potency (see section 4.6 for details). For comparison, we also conducted peptide binder design using RFd+MPNN and AfDesign. We used peptide ligand RMSD (L-RMSD) and peptide sequence similarity (calculated by Biopython's pairwise2 module[50]) to evaluate whether these models can generate ligands that resemble the original peptides in the complexes.

For these 30 nonredundant protein-peptide complexes, peptides generated by DiffPepBuilder showed a mean L-RMSD of 4.76 Å, while those of RFdiffusion and AfDesign were 13.62 and 10.76 Å, respectively. In terms of the average best L-RMSD (the minimum L-RMSD of all generated peptides for the same target), DiffPepBuilder achieved 3.34 Å, while those of RFdiffusion and AfDesign were 6.75 and 7.28 Å, respectively. DiffPepBuilder could generate peptide binders with similar sequence as the original binder peptides with the best average sequence similarity of 52.38% (ranging from 40.71% to 65.00%), while that of RFd+MPNN and AfDesign-designed sequences achieved average best similarity of 45.03% and 33.72%, respectively. The distribution of L-RMSD and similarity is illustrated in Figure 3a,b. These analyses demonstrate that DiffPepBuilder performs well in the regeneration test.

Figure 3c−e gives examples of regenerated peptide ligands with different types of secondary structures. Figure 3c displays an example of major histocompatibility complex class 2 (MHCII, PDB ID: 1SJH[51]), where our model generated an extended peptide that aligns well with the original backbone structure, achieving an L-RMSD of 2.42 Å and a sequence similarity of 44.4%. The second example corresponds to a peptide binder with a loop structure that binds human O-GlcNAc transferase (PDB ID: 6MA3[52]), where the DiffPepBuilder generated peptide has an L-RMSD of 1.49 Å and sequence similarity of 38.46%. Figure 3e illustrates an example that contains a peptide binder with helical structure, which is a complex structure of MDM2 (mouse double minute 2 homologue) with a 12-mer peptide inhibitor (PDB ID: 3EQS[53]), where DiffPepBuilder generated a helical peptide with an L-RMSD of 1.78 Å and a sequence similarity of 45.45%. MDM2 is an important regulator that inhibits tumor suppressor p53. Peptide inhibitors like ALRN-6924[54] have achieved good therapeutic effects in clinical settings. MDM2 peptide inhibitors share a relatively conserved sequence pattern, PxFxDYWxxL, where the hydrophobic residues F, W, and L play major roles in hydrophobic interactions with MDM2.[55,56] We further searched all generated sequences for MDM2 against this pattern and found 3 designs that recover all three of these critical conserved sites. We showcase one of these structures in Figure S2, from which we can see that the aforementioned conserved residues are closely aligned, demonstrating that our model holds the potential of generating novel inhibitors for MDM2.

In general, the best ddG (the lowest ddG of all the generated peptides on the same target) of peptide ligands generated by DiffPepBuilder is lower than those of the original peptides in the complexes, as shown in Figure 3f. We further analyzed the performance of DiffPepBuilder on peptide binders with different secondary structures. We differentiated helix from other conformations (mostly loop structures) by setting a criterion where the content of helix secondary structure in the ground truth peptide is at least 40%, with at least five consecutive residues adopting a helix conformation. As shown in Figure 3g,h, DiffPepBuilder performs better on helical structures for both sequence similarity and L-RMSD. In some cases, the loops generated by DiffPepBuilder do not match well with the native ligand, which may stem from the irregular nature of loop structures.

2.4. *De Novo* Peptide Binder Design Using DiffPepBuilder. We further used DiffPepBuilder to generate novel peptide binders for several important drug targets which do not have homologous proteins in the training data set to assess its capabilities of generating peptides from scratch and compared its performance to AfDesign and RFd+MPNN. We tested on three targets: $3CL^{pro}$, ALK1, and TNF-α. $3CL^{pro}$ is critical for SARS-CoV-2 viral maturation and replication, which can be inhibited by drugs like Pfizer's Paxlovid containing $3CL^{pro}$ inhibitor of Nirmatrelvir, yielding significant clinical benefits.[57] ALK1 is vital for angiogenesis regulation and serves as a key target in cancer therapy. Inhibiting the ALK1 signaling pathway that is crucial for tumor vasculature can substantially reduce tumor growth. Therapies targeting ALK1, such as the monoclonal antibody PF-03446962 and the trap receptor Dalantercept, disrupt its signaling to curb tumor progression.[58,59] TNF-α, a pivotal cytokine in immune regulation and inflammation, plays a significant role in chronic diseases, such as rheumatoid arthritis and psoriasis. Its central function in inflammation makes it a key target for drugs such as Infliximab and Adalimumab, which treat autoimmune disorders by neutralizing TNF-α to reduce symptoms and control inflammation.[60−62]

For $3CL^{pro}$, we selected the substrate pocket as the peptide targeting site (volume: 681.00 $Å^3$; surface area: 461.50 $Å^2$, as measured by CavityPlus[63,64]), which is suitable to accommodate extended peptides and would be too narrow for helical peptides. The crystal structure of $3CL^{pro}$ with a PDB ID of 7Z4S[65] was chosen to test the model's proficiency in generating nonhelical peptides. Conversely, the interface of ALK1 with BMP10 (Bone Morphogenetic Protein 10, PDB ID: 6SF1[66]) is more expansive, leading DiffPepBuilder to favor both helical and nonhelical peptides. As the interface features a significant proportion of hydrophobic residues (with a hydrophobic interface area of 958.46 $Å^2$ and a hydrophobic interface ratio of 67% as calculated by Rosetta[67]), the peptides need to cover more hydrophobic regions and balance the polar interactions for higher binding affinity. TNF-α forms a stable homotrimer, and its receptor TNFR1 binds to each of the grooves formed by two TNF-α protomers. This placement demands the generation of peptides capable of simultaneously interacting with two protein chains. Consequently, we selected TNF-α as a case study to evaluate the efficacy of DiffPepBuilder in addressing targets with binding sites formed by more than one chain. Specifically, we utilized the protein structure with PDB ID 7KP7,[68] which showcases the TNF-α binding interface with human TNFR1. Given the interfaces' large size (2843.77 $Å^2$), flatness, and hydrophilic nature, the design task needs to generate peptides with a large interaction surface and a sophisticated network of polar interactions.

For $3CL^{pro}$, as detailed in Table 1 and Figure 4a, DiffPepBuilder and RFd+MPNN generated potential peptide binders containing 8 to 30 residues with comparable ddG values. As peptides with about 10 residues are preferable for peptide drugs, we then tested the performance of different models with peptide length between 8 and 15 residues. DiffPepBuilder achieved the best performance, with an average ddG of −21.17 kcal/mol. In the case of ALK1, as shown in Figure 4c and Table 1, DiffPepBuilder achieved an average

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. *De novo* Generation Results[a]

Target	Method	Metric (Mean) ddG (kcal/mol, ↓)	ddG$_{8-15}$ (kcal/mol, ↓)	pTM-score (↓)	Validity % (↑)
3CL^{pro}	AfDesign	−12.33	−9.77	**0.22**	47.82
	RFd+MPNN	**−23.38**	−21.04	0.44	68.42
	DiffPepBuilder	−23.16	**−21.17**	0.27	**96.22**
ALK1	AfDesign	−7.61	−7.28	0.41	30.43
	RFd+MPNN	−20.31	−20.18	0.67	81.89
	DiffPepBuilder	**−24.62**	**−20.57**	**0.34**	**99.63**
TNF-α	AfDesign	−17.16	−19.98	0.30	71.36
	RFd+MPNN	−22.37	−16.26	0.54	92.76
	DiffPepBuilder	**−26.88**	**−23.63**	**0.20**	**96.83**

[a]The best in-group results are shown in **bold**.

Rosetta binding free energy of −24.49 kcal/mol, outperforming RFd+MPNN for peptide length of 8 to 30 residues. For peptides with 8 to 15 residues, DiffPepBuilder not only generated structures with the best binding energies (∼−45 kcal/mol) but also demonstrated better average values compared to RFd+MPNN. For TNF-α, as illustrated in Figure 4e and Table 1, DiffPepBuilder achieved the best average ddG of −26.88 kcal/mol, with the lowest value approaching about −60 kcal/mol. In generating peptides with 8 to 15 residues, DiffPepBuilder reached an average ddG of −23.63 kcal/mol, outperforming AfDesign and RFd+MPNN by a large margin. We also normalized the ddG over the number of residues, and DiffPepBuilder remains the best (Figure S5). Further interface analysis revealed that DiffPepBuilder excelled in minimizing buried unsatisfied hydrogen bonds, particularly for ALK1.

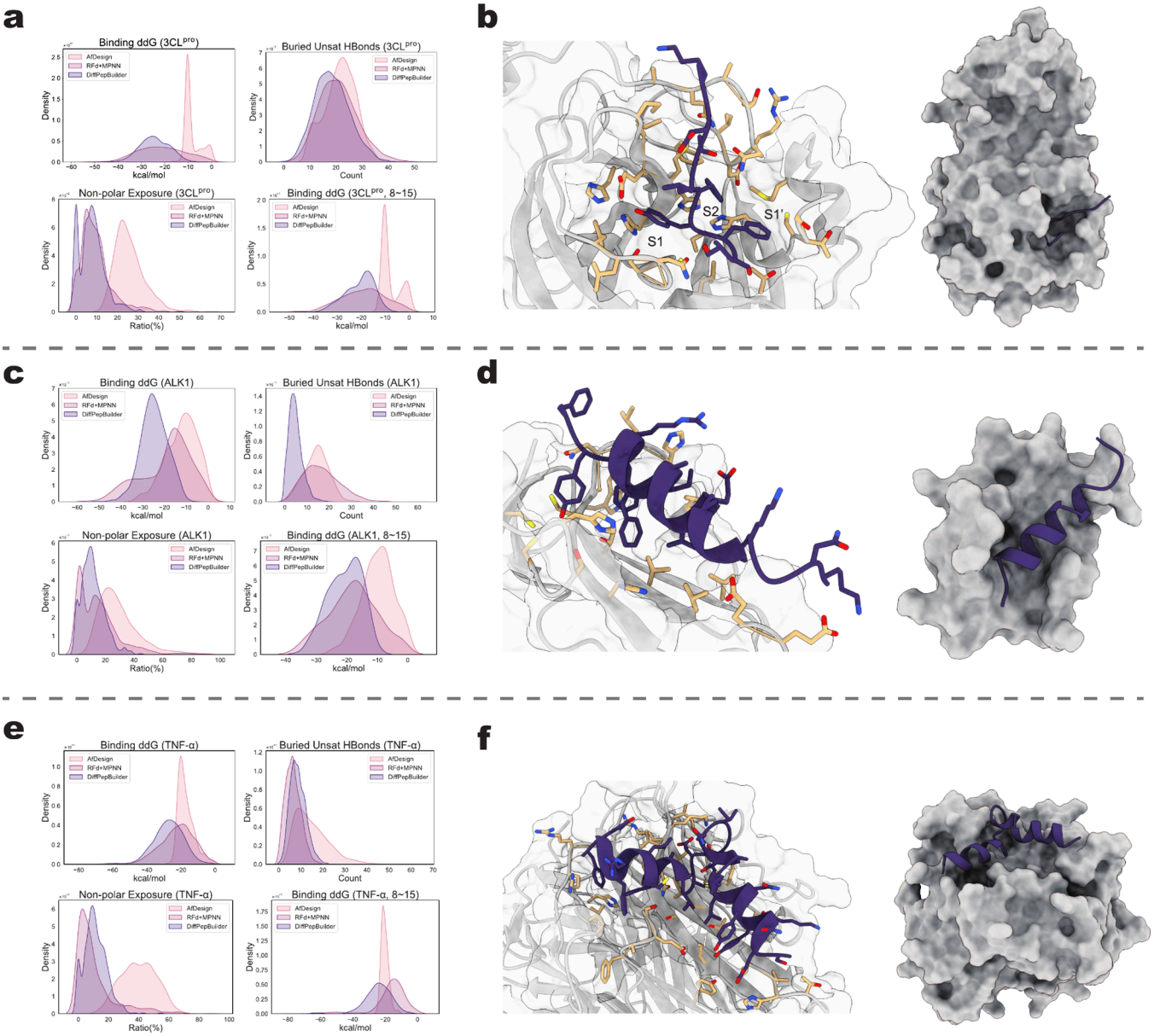


Figure 4. *De novo* generation performance of DiffPepBuilder. (a, c, and e) Distributions of interface metrics of binding ddG, buried unsatisfied hbond, nonpolar exposure ratio, and binding ddG of candidates with 8−15 residues targeting 3CL^{pro}, ALK1, and TNF-α, respectively. The performance of DiffPepBuilder is compared to that of AfDesign and RfD+MPNN. (b, d, and f) Examples of peptides generated by DiffPepBuilder (in violet) targeting three proteins (in gray, with interacting side chains in wheat): 3CL^{pro}, ALK1, and TNF-α, respectively.

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

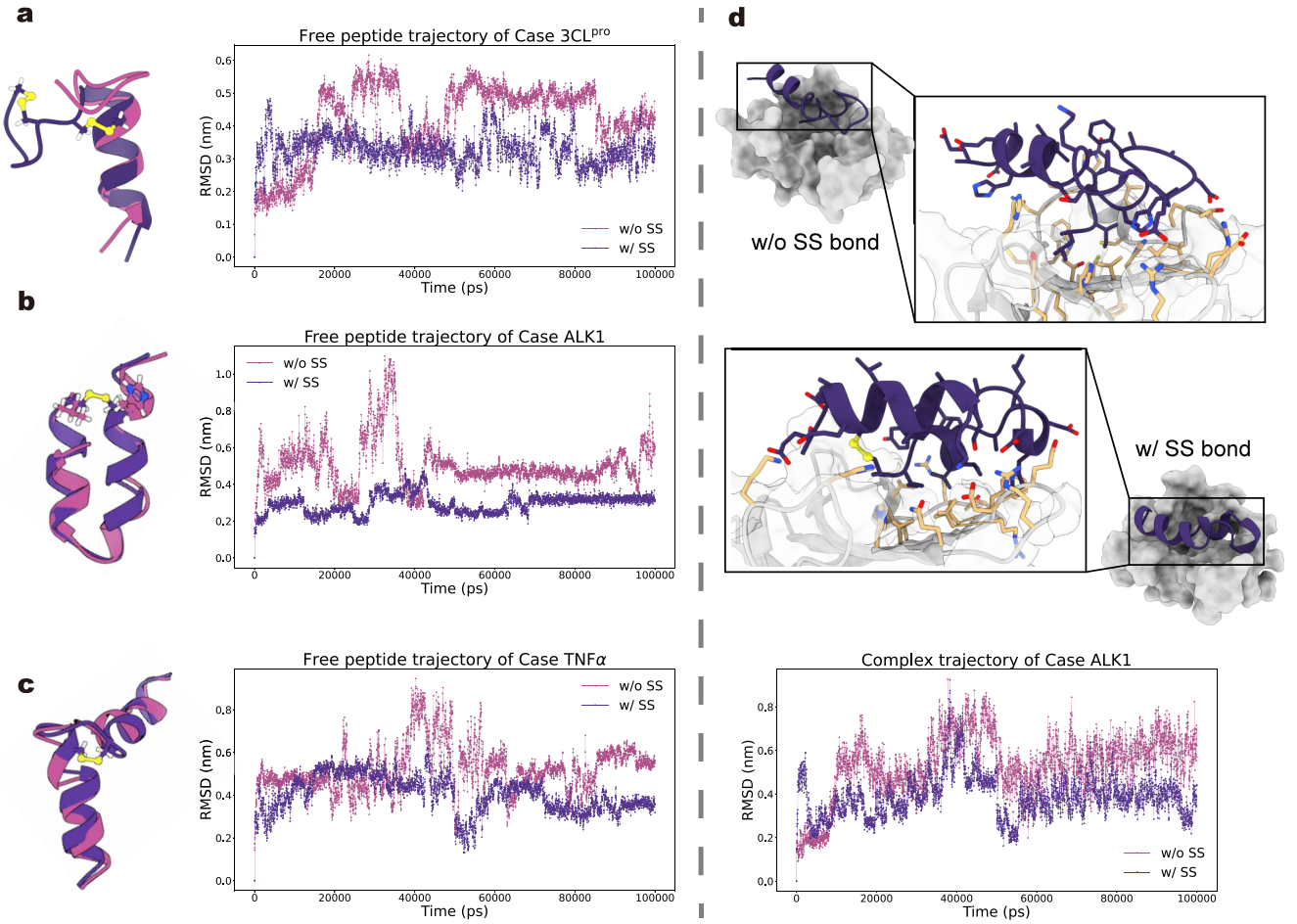


Figure 5. Case studies of the effect of disulfide bond construction. (a−c) Left: A comparison of free-state peptide structures generated by DiffPepBuilder for 3CL^{pro}, ALK1, and TNF-α, shown before (in pink) and after (in deep purple) disulfide bond construction, is shown at the end of MD simulations. Right: The RMSD trajectories from MD simulations. (d) Protein-peptide complex structures of ALK1 at the end of the MD simulations. The complexes before (above) and after (in the middle) disulfide bond construction are shown. The complex structure is illustrated in gray, with interacting side chains in wheat, highlighting both interfaces featuring hydrogen bonds and hydrophobic interactions. Down: The RMSD trajectories from MD simulations.

About 78.4% of its designs featured fewer than 10 unsaturated hydrogen bonds, with ≤20 hydrophobic exposure ratio mostly, as indicated in Figure 4a,c,e. Compared to DiffPepBuilder and RFd+MPNN, AfDesign-generated peptides suffer from low calculated binding free energy and a high ratio of hydrophobic exposure. We further measured the pTM-score (average TM-score of a generated peptide against all the remaining ones) as a metric of structure diversity. The mean pTM-scores of DiffPepBuilder for ALK1, 3CL^{pro}, and TNF-α were 0.34, 0.27, and 0.20, respectively, indicating a good diversity of the generated peptides. As shown in Table 1, the generation diversity of DiffPepBuilder is remarkably better than that of RFd+MPNN and comparable to that of AfDesign. Additionally, DiffPepBuilder achieved a higher validity (percentage of ddG < 0 for the generated peptides) than RFd+MPNN and AfDesign.

Figure 4b,d,f illustrates examples of generated structures for these three targets, respectively. In Figure 4b, the S1, S2, and S1′ subpockets of the 3CL^{pro} substrate binding pocket are occupied by the F5, Y6, and F7 residues of the generated peptide, forming complementary hydrophobic interactions, as well as hydrogen bonds. In Figure 4d, the interface of ALK1 is covered with the hydrophobic interactions by a generated helical peptide at the core region and forms hydrogen bonds with the ligand at the edge. Most residues presented outside the helix are polar or charged. In Figure 4f, the structure of a generated protein-peptide complex reveals that the peptide occupies the hydrophobic pockets on the groove of the TNF-α homotrimer interface and covers the hydrophobic areas of the β-sheet motif at the interface. It also forms relatively abundant polar interactions.

2.5. Stabilizing Peptide Binder Conformation by Introducing Disulfide Bond. We have shown that DiffPepBuilder performed well in the linear peptide binder design. However, linear peptides are often flexible in solution, which may result in large entropy loss upon binding to targets. They are also susceptible to protease degradation, which leads to short half-lives and weak efficacy in vivo.[7,69] In order to constrict the generated binding conformations and to enhance the binding potency and stability of the designed peptides, we introduced the SSBuilder module into DiffPepBuilder to design disulfide bonds in the generated peptides. After linear peptide generation for the aforementioned three cases, we further used DiffPepBuilder to generate disulfide-bond-containing peptides and conducted comparative analyses.

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

We first compared the binding strengths of the generated peptide binders with and without disulfide bonds and found that there is no obvious change in Rosetta ddG (see Figure S6). For each case study, we ranked the generated peptides by Rosetta binding energy and randomly selected 5 structures from the top 100 for MD simulation studies (details in section S3), including simulations of the free peptides and the protein-peptide complex. We performed a 100 ns MD simulation for each of the generated peptides in free form and calculated the average RMSD (compared to the designed conformation in the complex) of the last 20 ns of the trajectories. The results are shown in Figure 5a−c. Figure 5a presents an example for the $3CL^{pro}$ target, where SSbuilder constructs 2 disulfide bonds, one of which is located at the beginning and end of the peptide loop region, fixing the entire loop region into a cyclic structure, and the other is located at the hinge area where the cyclic loop connects with the helix. Figure 5b illustrates an example peptide generated for the ALK1 target, where the disulfide bond is positioned at the two terminal residues of the peptide, stabilizing the helical hairpin structure that reduced the RMSD from 4.5 to 3.1 Å. In the example of TNF-α (Figure 5c), the disulfide bond is also generated in the hinge area between two helical structures, anchoring the two helices together, which causes an RMSD decrease from 5.6 to 3.4 Å. We then filtered the peptides with minimal RMSD and perform 100 ns MD simulation on their complex structures. Figure 5d displays the structure and L-RMSD trajectory of ALK1 (Figure 5a). Structures without disulfide bonds exhibit a generally higher flexibility relative to that in the complex compared to those constrained by disulfide bonds. The peptide ligands with disulfide bonds tend to retain or form more stable hydrogen bonds and hydrophobic interactions. Subsequently, we selected the last 20 ns of the complex trajectory for MMPBSA analysis[70,71] and found that the peptide with disulfide bonds has not only a significantly lower calculated binding free energy but also smaller L-RMSD compared with the peptide without disulfide bonds. The other two candidate peptides with disulfied bonds of $3CL^{pro}$ and TNF-α showed a similar improvement on the binding conformation stability and binding free energy (Table 2). The candidate complex structures of ALK1 and $3CL^{pro}$ after MD are shown in Figure S8.

Table 2. Results of MD Simulations and MMPBSA Assays[a]

Case name	Metric		
	Free Peptide RMSD (Å)	Complex L-RMSD (Å)	ddG by MMPBSA (kcal/mol)
$3CL^{pro}$	4.8/3.2	7.8/3.2	−17.32/−47.66
ALK1	4.5/3.1	5.8/3.3	−30.98/−56.37
TNF-α	5.6/3.4	4.8/2.7	−34.45/−54.10

[a]The values preceding the "/" symbol represent those without disulfide bonds, whereas the values following are with disulfide bonds.

3. CONCLUSION AND DISCUSSION

Recent advancements in protein binder design have showed promising applications in biopharmaceutical and synthetic biology research and development, facilitated by improved computational methods, sophisticated structural biology techniques, and enhanced display technologies.[72,73] However, designing peptide binders remains challenging due to their lower stability, affinity, and specificity compared to proteins. The flexibility and rapid degradation of peptides add complexity to predicting and engineering effective peptide binders, and further innovations are needed to fully harness their potential, possibly involving non-natural amino acids or cyclization strategies to improve their biophysical properties.[74,75] In this study, we alleviated the scarcity of natural protein-peptide complex data by constructing a diverse, high-quality synthetic data set and developed DiffPepBuilder, a data-driven diffusion-based generative model. The results in the regeneration tests demonstrated that DiffPepBuilder is capable of generating peptides with conformations closely resembling natural ligands and possesses the potential to explore more optimal binding free energies. Compared with peptides with loop conformations, those with helical conformations can be regenerated better with similar sequences and backbones. This may be due to the inherently nonconservative nature of loop conformations, making it difficult for the model to learn a regular conformational distribution. In order to stabilize the binding conformations, DiffPepBuilder incorporates the SSBuilder module to construct disulfide bonds in the generated peptides. Molecular dynamics simulations and MMPBSA calculations demonstrated that integrating disulfide bonds into the generated peptide binders can indeed stabilize their binding conformations and enhance binding strength. We further tested the *de novo* peptide binder design ability of DiffPepBuilder on three key drug design targets. DiffPepBuilder outperforms the existing methods in performance, excelling in generating peptides with notable structural diversity and high affinity.

A major challenge in deep-learning-based, data-driven peptide binder design is the scarcity of high-affinity protein-peptide complex data. While data augmentation is a common practice in representation learning of protein−small molecule interactions,[76,77] to the best of our knowledge, similar work in peptide design is absent. In this study, we address the data scarcity issue by utilizing high-quality synthetic data derived from diverse protein−protein complex structures with a superior binding performance. DiffPepBuilder demonstrates the ability to capture essential interactions using these synthetic data and successfully regenerate native-like peptide binder structures. This alleviates concerns about a potentially significant gap between natural protein-peptide and synthetic protein-fragment complex data that could mislead the model. The superior *de novo* design performance of DiffPepBuilder on three real-world targets further validates our model's capability with data augmentation. Additionally, we acknowledge that the model would benefit from training on high-affinity natural protein-peptide data as its availability increases. We also envisage tuning our model on small-quantity data with high affinity via few-shot learning as a future direction. Moreover, the model's performance could be enhanced by employing a joint diffusion process for categorical and continuous variables, a method that has demonstrated efficacy in the generation of small molecules[78,79] and proteins.[80,81] Additionally, recent advancements in flow matching[82,83] offer a compelling alternative to diffusion models, presenting a promising avenue for future exploration.

As a target-specific *de novo* peptide design method, DiffPepBuilder requires target protein structures, which may be either experimentally determined or *in silico* predicted, along with prior binding site information, as input. In cases where native binding peptides or molecules are already identified, binding site information can be determined either

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

automatically through DiffPepBuilder's Reference Ligand mode or by manually analyzing the interface interactions. Conversely, in the absence of prior binding site data, established pocket searching algorithms such as CavityPlus[63,64] can be incorporated into the DiffPepBuilder workflow to identify potential binding pockets of target proteins. These pockets are further analyzed to identify hotspot residues via pharmacophore analysis.[84] Currently, DiffPepBuilder does not take the flexibility of the input target structures into account during model inference. This poses a challenge as conformational changes are frequently observed during the peptide binding process,[85−88] and relying on the rigid structure of target proteins may decrease the model's performance in generating optimal binders. We are actively addressing this challenge by integrating receptor flexibility into DiffPepBuilder to further enhance its generative performance.

Turning linear peptide molecules into cyclic ones is an important strategy in peptide drug development.[69] Cyclization by introducing disulfide bonds could reduce flexibility and decrease the entropy loss upon target binding,[89−91] serving as an important method to enhance the stability of peptide drugs.[92] Currently DiffPepBuilder only includes disulfide bonds as conformation restriction methods; in the future, we will further introduce more diverse cyclization methods either with natural or unnatural peptide cyclilization methods. Another issue in peptide drug design is the relatively small sequence space of peptides compared with the chemical space of small molecules or the sequence space of proteins. Further exploration of chemical modification in peptide structures based on increasing experimental data can provide guidance for the design of peptides with unnatural residues, which, incorporating data-driven computational methods, can further broaden the scope of peptide design.

4. METHODS

4.1. Data Set Construction. *Structure Collection.* Structures were selected from the Protein Data Bank prior to October 2023, based on the following criteria: composed solely of proteins, comprising more than two chains, and having a resolution lower than 2.5 Å. Symmetric structures were complemented using biounit information, and missing atoms were rectified with PDBFixer.[93] A total of 5,358 structures with the shortest chain length between 8 and 30 residues were collected as the peptide−protein data set, while 11,469 structures with the shortest chain length greater than 30 residues were collected as the protein−protein data set. For the latter group, the shortest chain of the complex was defined as the ligand, and any chains interacting with the ligand, defined by an interface area (dSASA) of ≥100 Å^2, were output in pairs. The permanent protein−protein complexes, which have a buried surface area larger than 3,000 Å^2, were not excluded, resulting in 35,338 protein dimer structures.

Interfacial Fragment Extraction. For pairwise protein−protein complexes, we first conducted an analysis on the solvent accessible surface area (SASA) of both the ligand and the complex, utilizing the BioPython software package.[50] Then we defined residues with a buried surface area (BSA, $SASA_{ligand} - SASA_{complex}$) greater than 36 Å^2 and a buried proportion (computed by $(SASA_{ligand} - SASA_{complex})/SASA_{ligand}$) exceeding 40% as buried residues. Within the range of 8−30 residues, we established a series of truncation windows. For each chain in a dimer, we slid these windows across, identifying a segment as a helical structure if it contains a continuous sequence of ≥8 residues in a helical conformation. The others are nonhelical. For helical segments, the proportion of hotspot residues is ≥40%; for nonhelical ones, the requirement was ≥80%. For each complex structure, we conducted an initial deduplication check before the overall sequence redundancy reduction, removing redundant entities from the set of fragments associated with the same dimer complex, which was done according to the criteria of L-RMSD ≤ 0.5 Å (same length) and ligand sequence similarity ≥70%. This process resulted in 34,618 entries.

Redundancy Reduction. Redundancy reduction was implemented on both peptide−protein and protein−protein data sets. It was carried out from two aspects: sequence similarity and structural similarity. The measurement of sequence similarity involved conducting a multisequence alignment of the longest chain by CD-HIT[94−96] with a 90% similarity cutoff. Then, in each cluster, the structure with the highest X-ray resolution was chosen as the reference, and then the L-RMSD of Cα atoms and sequence similarity of other ligands (the shortest chain) in the cluster were calculated. Structures with L-RMSD larger than 5 Å and similarity lower than 70% were preserved for further analysis. This redundancy reduction was implemented for both PepPC (peptide−protein complex) and PepPC-F (peptide−protein complex fragments). After the redundancy reduction process, 117 clusters were generated from 275 helical-ligand complexes and 1,021 clusters of nonhelical-ligand complexes in PepPC; 2,588 clusters were generated from 16,553 helical-ligand complexes and 3,386 clusters of 18,065 loop-ligand complexes in PepPC-F. After further data set cleaning, including exclusion of disulfide bond and membrane protein (see the SI for details), there were 3,832 structures (232 helical ligands) in PepPC and 14,897 structures (4,241 helical ligands) in PepPC-F.

Secondary Structure. We calculated the secondary structure information for all peptide ligands using the DSSP module in Biopython.[50] In this context, Helix includes A-Helix, 5-Helix, and 3-Helix; Beta includes B-sheet; Turn includes B-Bridge, Bend, and Turn; Unstructured represents Coil.

4.2. Input Preparation. In the peptide binder *de novo* generation process, DiffPepBuilder initially takes the full target as input and truncates it to the binding pocket according to the user-provided binding site information. Specifically, the binding site information can be formatted into three types, each corresponding to one of the three pattern diagrams shown in Figure 1a. The first type, Motif, requires the user to provide information about the target binding motif (can be a structurally truncated PDB file or sequence information). The second type, Hotspots, requires hotspot information (often provided as strings of hotspot residue IDs). The third type, Reference Ligand, allows for specifying a reference ligand by inputting a PDB file and specifying the receptor. The model will delineate residues on the receptor that are a certain distance from the ligand Cα (default is 8 Å) as the hotspots. We tested several cutoff values and found that the 8 Å cutoff maintains an optimal protein motif size (see Figure S4 and Table S2). For all three types of information, residues whose Cα atom are within 10 Å of the Cα atoms from these binding motif or hotspots compose the binding pocket. The model then calculates the geometric center of the pocket's Cα atoms and adds a σ_0 = 2 Å Gaussian noise to the center position to enhance the generation diversity:

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$\boldsymbol{x}_{\text{center}} = \sum_{i=1}^{N_{\text{rec}}} \boldsymbol{x}_i^{\text{rec}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_0 \boldsymbol{I}) \tag{1}$$

where $\boldsymbol{x}_i^{\text{rec}}$ represents the coordinates of the Cα atom of the i-th pocket residue and N_{rec} is the number of pocket residues. For each peptide length N_{lig} within the user-defined range $[N_{\min}, N_{\max}]$, the model independently samples N_{lig} positions, $\{\boldsymbol{x}_i^{\text{lig}}\}_{i=1}^{N_{\text{lig}}}$, from an isotropic Gaussian distribution $\mathcal{N}(\boldsymbol{x}_{\text{center}}, \sigma \boldsymbol{I})$, with σ acting as an adjustable hyper-parameter that controls the structural diversity of the *de novo* generated peptides. The model subsequently recenters the target pocket-peptide ligand complex to the noise-adjusted centroid $\boldsymbol{x}_{\text{center}}$ to ensure the (approximate) SE(3)-equivariance in the diffusion process, aligning with methodologies suggested by previous works:[46,97,98]

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \boldsymbol{x}_{\text{center}} \tag{2}$$

here $\boldsymbol{x}_i$ denotes the Cα coordinates of an arbitrary residue i within the pocket-peptide ligand complex. Following this, the model constructs a fully connected graph of the complex, which is then processed through the diffusion-based generative procedure. The model additionally generates a binary mask, $\boldsymbol{m} \in \{0, 1\}^N$, that indicates the positions of the pocket residues that are to be fixed during the diffusion process.

Prior to the diffusion-based generative procedure, the backbone of the target pocket-peptide ligand complex is parametrized as a collection of rigid-body frames, $\mathbf{T} = [\boldsymbol{T}_1, ..., \boldsymbol{T}_N]$, akin to the approach in AlphaFold 2,[28] where $N = N_{\text{rec}} + N_{\text{lig}}$. Note that each frame $\boldsymbol{T}_i = (\boldsymbol{R}_i, \boldsymbol{x}_i)$ consists of a 3-dimensional rotation $\boldsymbol{R}_i \in \text{SO}(3)$ and a translation $\boldsymbol{x}_n \in \mathcal{R}^3$. Initial frames of the peptide to be generated are independently sampled from a uniform distribution on SO(3), i.e., $\mathcal{U}^{\text{SO}(3)}$.

4.3. Architecture of the Denoising Network. The diffusion model utilized in DiffPepBuilder essentially introduces multilevel noises to the ground-truth peptide ligand conformation and learns to recover the original structure through a denoising network. During the inference process, it reverses this noising process and *de novo* generates the peptide ligand conditioned on the pocket (Figure 1b). Mathematically, the forward noising process and the reverse denoising process on $\text{SE}(3)^N$ are characterized by

$$\text{d}\mathbf{T}^{(t)} = \left[0, -\frac{1}{2}\text{P}\mathbf{X}^{(t)}\right]\text{d}t + [\text{d}\mathbf{B}^{(t)}_{\text{SO}(3)^N}, \text{d}\mathbf{B}^{(t)}_{\mathcal{R}^{3N}}] \tag{3}$$

and

$$\text{d}\overleftarrow{\mathbf{R}}^{(t)} = \nabla_R \log p_t(\overleftarrow{\mathbf{T}}^{(t)})\text{d}t + \text{d}\mathbf{B}^{(t)}_{\text{SO}(3)^N} \tag{4}$$

$$\text{d}\overleftarrow{\mathbf{X}}^{(t)} = \text{P}\left(\frac{1}{2}\mathbf{X}^{(t)} + \nabla_x \log p_t(\overleftarrow{\mathbf{T}}^{(t)})\right)\text{d}t + \text{P}\text{d}\mathbf{B}^{(t)}_{\mathcal{R}^{3N}} \tag{5}$$

respectively,[46,99] where P is the projection matrix removing center of mass. Yim et al.[46] proved that the score, i.e., $[\nabla_R \log p_t, \nabla_x \log p_t]$, can be computed separately for the rotation and translation of each residue. The denoising network is trained to approximate the score with $[s^R_\theta, s^x_\theta]$, a process known as denoising score matching (DSM).[43,44]

As depicted in Figure 1d, the denoising network of DiffPepBuilder is composed of three main components: the Embedding Module, the Cross Update Module, and the Multi-Task Decoder. In the **embedding stage**, initial node embeddings $\mathbf{h}_0 = [h_1, ..., h_N]$ are generated by concatenating diffusion time step embeddings, residue index embeddings, residue type embeddings, and pLM embeddings. We employed sinusoidal positional embeddings[100] for residue index embeddings and utilized a frozen 650 M parameter ESM-2[101] encoder to extract the pLM embeddings from the sequence of the untruncated target. Peptide ligand residues are assigned a special residue type `<mask>` in the residue type embedding process, and the average pLM embeddings of pocket residues are used as their pLM embeddings. To indicate breaks between chains, we incorporated a 100-residue gap in the residue indices between each chain, following the approach in RoseTTAFold.[29] Node embeddings $\mathbf{h}_0$ are subject to cross-concatenation and are subsequently concatenated with self-conditioning distogram embeddings[35,102] and sequence distance embeddings, resulting in the preliminary edge embeddings $\mathbf{z}_0 = (z_{ij})_{i,j=1}^N$. Following RFdiffusion, we performed self-conditioning of the predicted Cα pairwise distances, $\hat{\mathbf{d}}$, in 50% of the examples in the training process.

The initial node embeddings $\mathbf{h}_0$, edge embeddings $\mathbf{z}_0$, and backbone frames $\mathbf{T}_0$ are subsequently processed through the **Cross Update Module** (Figure S3). Specifically, within an arbitrary layer l, the node update is achieved through a combination of Invariant Point Attention (IPA),[28] Transformer encoder layers,[100] Multi-Layer Perceptron (MLP), Layer Normalization (LayerNorm),[103] and Linear layers:

$$\mathbf{h}_{\text{ipa}} = \text{LayerNorm}(\text{IPA}(\mathbf{h}_l, \mathbf{z}_l, \mathbf{T}_l) + \mathbf{h}_l) \tag{6}$$

$$\mathbf{h}_{\text{trans}} = \text{Transformer}(\mathbf{h}_{\text{ipa}} \| \text{Linear}(\mathbf{h}_0)) \tag{7}$$

$$\mathbf{h}_{l+1} = \text{MLP}(\mathbf{h}_{\text{ipa}} + \text{Linear}(\mathbf{h}_{\text{trans}})) \tag{8}$$

here $\|$ denotes the concatenation process. Subsequently, the edge update is performed as

$$\mathbf{h}_{\text{proj}} = \text{Linear}(\mathbf{h}_{l+1}) \tag{9}$$

$$\mathbf{z}_{\text{concat}} = \mathbf{z}_l \| \text{CrossConcat}(\mathbf{h}_{\text{proj}}) \tag{10}$$

$$\mathbf{z}_{l+1} = \text{LayerNorm}(\text{MLP}(\mathbf{z}_{\text{concat}})) \tag{11}$$

here `proj` is an abbreviation for projection, which is implemented by a Linear layer. Lastly, the frame update is executed following the `BackboneUpdate` algorithm in AlphaFold 2:

$$\mathbf{T}_{\text{update}} = \text{CalcRot}(\text{Linear}(\mathbf{h}_l)) \tag{12}$$

$$\mathbf{T}_{\text{masked}} = \text{MaskRec}(\mathbf{T}_{\text{update}}, \boldsymbol{m}) \tag{13}$$

$$\mathbf{T}_{l+1} = \mathbf{T}_l \cdot \mathbf{T}_{\text{masked}} \tag{14}$$

where the `CalcRot` function transforms the input nonunit quaternion into a rotation matrix $\boldsymbol{R}_{i,\text{update}}$ and outputs it along with the coordinate update $\boldsymbol{x}_{i,\text{update}}$. The `MaskRec` function modifies the frame update of target pocket residues to $\boldsymbol{T}_i^{\text{rec}} = (\boldsymbol{R}_i^{\text{rec}}, \boldsymbol{x}_i^{\text{rec}}) = (\boldsymbol{I}, \boldsymbol{0})$, ensuring that the target remains fixed during the diffusion process. We stacked $L = 6$ layers with no shared parameters to form the Cross Update Module.

The updated node embeddings $\mathbf{h}_L$, edge embeddings $\mathbf{z}_L$, and backbone frames $\mathbf{T}_L$ are then fed into the **Multi-Task Decoder**. This decoder predicts the translational and rotational scores of the current diffusion step, the residue types of the peptide ligand, the main chain torsion angles ψ, and the side chain torsion angles χ_1–χ_4. We take the final frame

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

representation $\mathbf{T}_L \equiv \hat{\mathbf{T}}^{(0)} = (\hat{\mathbf{R}}^{(0)}, \hat{\mathbf{X}}^{(0)})$ to calculate the translational score s^x_θ and rotational score s^x_θ following Yim et al.[46] Torsion angle prediction is implemented as (using ψ prediction as an example)

$$\mathbf{h}_{\text{psi}} = \text{MLP}(\mathbf{h}_L) \tag{15}$$

$$\boldsymbol{\psi}_{\text{unnorm}} = \text{Linear}(\mathbf{h}_{\text{psi}} + \mathbf{h}_L) \tag{16}$$

$$\boldsymbol{\psi}_{\text{pred}} = \text{Normalize}(\boldsymbol{\psi}_{\text{unnorm}}) \tag{17}$$

where the `Normalize` function is used to normalize the output torsion angles to be −180° to 180°. Residue type prediction is achieved in a one-shot decoding manner:

$$\mathbf{d}_{\text{pred}} = \text{CalcPairDist}(\mathbf{X}_L) \tag{18}$$

$$\mathbf{z}_{\text{mean}} = \text{RowMean}(\text{DistMask}(\mathbf{z}_L, \mathbf{d}_{\text{pred}})) \tag{19}$$

$$\mathbf{z}_{\text{logits}} = \text{MLP}(\mathbf{h}_L \| \mathbf{z}_{\text{mean}}) \tag{20}$$

$$\mathbf{s}_{\text{prob}} = \text{SoftMax}(\mathbf{z}_{\text{logits}}/\tau) \tag{21}$$

$$\boldsymbol{s}_{\text{pred}} = \text{Sample}(\mathbf{s}_{\text{prob}}) \tag{22}$$

where the `CalcPairDist` function calculates pairwise distances of Cα atoms based on given Cα coordinates. The `DistMask` function masks residue pairs whose Cα distances exceed a specified threshold, $d_{\max}$ = 12 Å. The sampling temperature τ is a hyper-parameter that controls the randomness of the sampling process. We set τ to 0.1 throughout the generation process. The `Sample` function samples from multinomial distribution $\mathbf{s}_{\text{prob}}$ to determine the final amino acid type $\boldsymbol{s}_{\text{pred}}$. In the sampling process, the peptide backbone frames are first iteratively denoised, and the residue types are subsequently predicted by leverageing the final (t = 0) backbone frames. During the sampling of the residue types, we additionally calculate per-residue entropy as

$$S_i = -\sum_{k=1}^{20} s_{\text{prob},k} \log s_{\text{prob},k} \tag{23}$$

where $s_{\text{prob},k}$ represents the probability of the k-th amino acid type for residue i of an arbitrary peptide ligand.

After the denoising process, DiffPepBuilder reconstructs the backbone of the generated peptide ligand using the final frame representations and the predicted ψ torsion angles as that in AlphaFold 2[28] and ultimately rebuilds the side chain conformations with PDBFixer[93] according to the predicted Cβ orientation to obtain an all-atom structure. The predicted complex structures undergo optimization using Rosetta's FastRelax[104] with a Fixbb (Fixed backbone) protocol to eliminate clashes of the side chains.

4.4. Details of SSBuilder. The **SSBuilder** method is based on a structure library of disulfide-bonded pairs, which is extracted from the PDB entries with an X-ray resolution <2.5 Å and prior to December 2023. We extracted all the structures of disulfide-bonded two cysteine residues. We then calculated the values of the dihedral angles formed by Cα1-Cβ1-Cβ2-Cα2 (Dihedral), the values of the two bond angles Cα1-Cβ1-Cβ2 (angle1) and Cβ1-Cβ2-Cα2 (angle2), as well as the value of the Cβ1-Cβ2 (distance), to serve as geometric matching criteria (see Figure S9). Angles were binned every 5°, and distance bins' widths were set to 0.1 Å.

Residue entropy in our deep learning model output was used to identify the potential disulfide bond site. This serves as an additional filter, excluding residues critical to binding interactions based on the model's confidence in residue type, thereby ensuring that only nonessential residues are considered for cysteine substitution. Residues with an entropy greater than 0.01 will be selected and paired. The angles and distance parameters of each residue pair are quickly matched to corresponding bins (those corresponding bins that cannot be found will be skipped). This is followed by a comparison with the same geometric parameters of the disulfide bond building blocks within those bins. The residue pair that achieves the best parameter matching (where the sum of the absolute differences of all parameters is minimized) is then spliced with its corresponding disulfide bond unit to create a cyclized structure.

4.5. Training Details. In the training of the denoising network, we employed a composite loss scheme that includes denoising score matching (DSM) loss $\mathcal{L}_{\text{dsm}}$, peptide amino acid type loss $\mathcal{L}_{\text{aa}}$, and a series of auxiliary losses comprising peptide backbone position loss $\mathcal{L}_{\text{bb}}$, pairwise atomic distance loss $\mathcal{L}_{\text{dist}}$, side chain torsion angle loss $\mathcal{L}_{\text{chi}}$, and Cα atom clash loss $\mathcal{L}_{\text{clash}}$. The SE(3) DSM loss is given by

$$\begin{aligned}\mathcal{L}_{\text{dsm}} = {} & \mathbb{E}_{i,t}[\lambda_t^x \| \nabla_x \log p_{t|0}(\boldsymbol{x}_i^{(t)} | \hat{\boldsymbol{x}}_i^{(0)}) - s^x_\theta(t, \mathbf{T}_i^{(t)}) \|] \\ & + \mathbb{E}_{i,t}[\lambda_t^R \| \nabla_R \log p_{t|0}(\boldsymbol{R}_i^{(t)} | \hat{\boldsymbol{R}}_i^{(0)}) - s^R_\theta(t, \mathbf{T}_i^{(t)}) \|]\end{aligned} \tag{24}$$

where $i \in \{1, \ldots, N_{\text{lig}}\}$, $t \sim \mathcal{U}[0, 1]$. We utilized the weight schedule following previous works:[44,46]

$$\lambda_t^R = 1/\mathbb{E}([\|\nabla \log p_{t|0}(\boldsymbol{R}_i^{(t)} | \boldsymbol{R}_i^{(0)})\|^2_{\text{SO}(3)}]) \tag{25}$$

$$\lambda_t^x = e^{t/2}(1 - e^{-t}) \tag{26}$$

The peptide amino acid type loss $\mathcal{L}_{\text{aa}}$ is calculated as a residue-averaged cross entropy loss for all peptide residues. For the auxiliary losses, the peptide backbone position loss is formulated as a mean squared error (MSE) loss on backbone atom positions:

$$\mathcal{L}_{\text{bb}} = \frac{1}{4N_{\text{lig}}} \sum_{i=1}^{N_{\text{lig}}} \sum_{n} \left\| \boldsymbol{x}_{i,n}^{(0)} - \hat{\boldsymbol{x}}_{i,n}^{(0)} \right\|^2 \tag{27}$$

where $n \in \{\text{N}, \text{C}, \text{C}\alpha, \text{O}\}$. The pairwise atomic distance loss is also an MSE loss on pairwise atomic distances:

$$\mathcal{L}_{\text{dist}} = \frac{1}{Z_{\text{dist}}} \sum_{i=1}^{N} \sum_{j=1}^{N_{\text{lig}}} \sum_{n,m} \mathbb{1}(d_{ij}^{nm} < d_{\text{dist}}) \| d_{ij}^{nm} - \hat{d}_{ij}^{nm} \|^2 \tag{28}$$

where d_{dist} = 6 Å, $n, m \in \{\text{N}, \text{C}, \text{C}\alpha, \text{O}\}$, and the normalizing factor Z_{dist} is given by

$$Z_{\text{dist}} = \sum_{i=1}^{N} \sum_{j=1}^{N_{\text{lig}}} \sum_{n,m} \mathbb{1}(d_{ij}^{nm} < d_{\text{dist}}) - N_{\text{lig}} \tag{29}$$

The side chain torsion angle loss is an MSE loss as computed in AlphaFold 2.[28] We only considered χ_1 and χ_2 angles because χ_3 and χ_4 are relatively less informative and more challenging to predict accurately.[105] We empirically found that this side chain torsion angle loss helps the model better capture inter-

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

residue interactions. The Cα atom clash loss $\mathcal{L}_{\text{clash}}$ is defined as

$$\mathcal{L}_{\text{clash}} = \frac{1e3}{Z_{\text{clash}}} \sum_{i=1}^{N_{\text{rec}}} \sum_{i=1}^{N_{\text{lig}}} \mathbb{I}(d_{ij}^{C\alpha} < d_{\text{clash}}) \tag{30}$$

where we set the clash threshold d_{clash} = 4 Å, and the normalizing factor Z_{clash} is computed by $Z_{\text{clash}} = N_{\text{rec}} \times N_{\text{lig}}$. The C$\alpha$ atom clash loss is utilized to minimize steric clashes between the target and generated peptide ligand. The full training loss is formulated as

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{dsm}} &+ \mathbb{I}(t < 0.25) w_1 \mathcal{L}_{\text{aa}} \\ &+ \mathbb{I}(t < 0.25) w_2 (\mathcal{L}_{\text{bb}} + \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{chi}}) \\ &+ \mathbb{I}(t < 0.5) w_3 \mathcal{L}_{\text{clash}}\end{aligned} \tag{31}$$

where we set the loss weights $w_1 = 2$, $w_2 = 0.25$, $w_3 = 0.25$. We applied the amino acid type loss $\mathcal{L}_{\text{aa}}$ and auxiliary losses primarily near $t = 0$ to encourage the model to learn fine-grained characteristics.

The denoising network consists of ∼104 M parameters and was trained exclusively on the PepPC-F data set using 8 NVIDIA A800 80GB GPUs. The training lasted ∼5 days. We employed the AdamW optimizer[106] with a learning rate of 1e-5. For multi-GPU training and inference, we used `DistributedDataParallel` implemented by PyTorch.[107]

4.6. Regeneration Details. For the regeneration test set PepPC-HF, we first collected a total of 822 structures from the PepPC data set that overlapped with the PDBbind2020 database.[47] Redundant entries were removed with a 40% sequence similarity threshold using CD-HIT.[94−96] We then filtered out entries with an activity value (K_i, K_d, IC_{50}) ≤ 0.1 μM and conducted further deduplication against the PepPC-F data using a maximum 60% sequence similarity criterion for target proteins (mostly ≤30%). Ultimately, these processes yielded 30 complex structures featuring high-activity peptide ligands (see Table S1 for details).

For **DiffPepBuilder**, we used the Reference Ligand mode and set the peptide length to that of the original peptide. We set the number of denoising steps, the noise scale, and the sampling temperature to 500, 1.0, and 0.1, respectively. For **AfDesign**, we utilized its fine-tuned version for generating peptide binders.[45] We set the weights for various parameters as follows: solubility at 0.5, msa_ent (multiple sequence alignment entropy) at 0.01, pLDDT (predicted Local Distance Difference Test) at 0.1, pae_intra (predicted aligned error for intrachain) at 0.1, and pae_inter (predicted aligned error for interchain) at 1.0. The weights for con_intra (contact predictions within chains) and con_inter (contact predictions between chains) were set to 0.1 and 0.5, respectively. The cycles for the three stages were 100, 100, and 10. For **RFdiffusion** used for peptide backbone generation, we set noise scale to 0. We employed the dlbinder design method[108] for the subsequent ProteinMPNN[34] sequence design and Rosetta side-chain assembly and FastRelax. Pocket residues within 5 Å of the reference peptide were selected as the hotspots for all three methods, and each method performed 128 samplings.

4.7. *De Novo* Generation Settings. For **DiffPepBuilder**, we sampled 128 times for each peptide length. The number of denoising steps is set to 500. We set the noise scale to 1.0, 0.5, and 2.0 for ALK1, $3CL^{pro}$, and TNF-α, respectively. The impact of the noise scale settings is illustrated in Figure S7. We note that adjusting the noise scale provides versatile control over the binding strength and generation diversity. For **AfDesign**, we adopted the parameters in section 4.6 and executed 128 sampling for each length. For **RFdiffusion**, we adhered to the default noise scale settings and employed the dlbinder design method as shown in section 4.6. We sampled 128 times for each specified length. For sequence design, we generated one sequence for each backbone. Peptide length was set to 8−30 for generation for each method, and the structures generated by all programs were optimized using Rosetta FastRelax[104] before evaluation. The interface parameters, including ddG and buried unsatisfied hydrogen bonds, were also calculated by Rosetta.

As there are no reference peptide ligands available in de novo generation studies, we used hot spots to define peptide binding sites. Hotspots for $3CL^{pro}$, ALK1, and TNF-α are "B24-B45-B46-B49-B140-B142-B143-B144-B145-B163-B164-B165-B166-B168-B188-B189-B191-B192", "B40-B58-B59-B71-B72-B87", and "C23-C25-C138-C139-C140-C141-D67-D68-D69-D71-D73-D79-D80-D81-D82-D83-D84-D87-D88-D89-D125-D127-D128-D129", respectively. For $3CL^{pro}$, based on complex structure in 7Z4S, we initially selected all residues on the $3CL^{pro}$ receptor that are within 8 Å of the Cα atoms of the cyclic peptide residues. We then performed an alanine scan using Rosetta on these residues. Those with $ddG_{mutant} - ddG_{wild\text{-}type} \geq 1$ kcal/mol are identified as hotspots. For ALK1, in the protein−protein complex structure of 6SF1, we used BMP10 as a reference and selected hotspots with the same criteria of $3CL^{pro}$. For the TNF-α, one of three TNFR1s in 7KP7 was selected as reference ligand, hotspots were selected from the TNF-α trimer structure according to the same criteria. All models use the same hotspot information as input for peptide generation.

Code Availability. The source code is available at https://github.com/YuzheWangPKU/DiffPepBuilder.

Data Availability. The PepPC and PepPC-F data sets are available at https://github.com/YuzheWangPKU/DiffPepBuilder/tree/main/datasets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00975.

Detailed descriptions of data set processing and division, specifics of the regeneration task, illustration of the model's cross update module, *de novo* generation task, molecular dynamics simulation parameter settings, and statistical details of SSbuilder parameters (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Changsheng Zhang − *BNLMS, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China;* orcid.org/0000-0002-8990-0878; Email: changshengzhang@pku.edu.cn

Luhua Lai − *Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China; BNLMS, College of Chemistry and Molecular Engineering and Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University,*

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Beijing 100871, China; orcid.org/0000-0002-8343-7587; Email: lhlai@pku.edu.cn

Authors

Fanhao Wang − *Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China;* orcid.org/0009-0006-6683-4720

Yuzhe Wang − *Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China;* orcid.org/0009-0007-3680-0308

Laiyi Feng − *Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00975

Author Contributions

†F.W. and Y.W. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (2022YFA1303700), the National Natural Science Foundation of China (21977007, T2321001), and the Chinese Academy of Medical Science (2021-I2M-5-014).

■ REFERENCES

(1) Bodanszky, M. *Peptide chemistry. A Practical Textbook*; 1988.

(2) Craik, D. J.; Fairlie, D. P.; Liras, S.; Price, D. The future of peptide-based drugs. *Chemical biology & drug design* **2013**, *81*, 136−147.

(3) Fosgerau, K.; Hoffmann, T. Peptide therapeutics: current status and future directions. *Drug discovery today* **2015**, *20*, 122−128.

(4) Gomes, B.; Augusto, M. T.; Felício, M. R.; Hollmann, A.; Franco, O. L.; Gonçalves, S.; Santos, N. C. Designing improved active peptides for therapeutic approaches against infectious diseases. *Biotechnology advances* **2018**, *36*, 415−429.

(5) Muttenthaler, M.; King, G. F.; Adams, D. J.; Alewood, P. F. Trends in peptide drug discovery. *Nat. Rev. Drug Discovery* **2021**, *20*, 309−325.

(6) Kaspar, A. A.; Reichert, J. M. Future directions for peptide therapeutics development. *Drug discovery today* **2013**, *18*, 807−817.

(7) Henninot, A.; Collins, J. C.; Nuss, J. M. The current state of peptide drug discovery: back to the future? *Journal of medicinal chemistry* **2018**, *61*, 1382−1414.

(8) Lee, A. C.-L.; Harris, J. L.; Khanna, K. K.; Hong, J.-H. A comprehensive review on current advances in peptide drug development and design. *International journal of molecular sciences* **2019**, *20*, 2383.

(9) Wang, L.; Wang, N.; Zhang, W.; Cheng, X.; Yan, Z.; Shao, G.; Wang, X.; Wang, R.; Fu, C. Therapeutic peptides: current applications and future directions. *Signal Transduction and Targeted Therapy* **2022**, *7*, 48.

(10) Chen, Z.; Wang, R.; Guo, J.; Wang, X. The role and future prospects of artificial intelligence algorithms in peptide drug development. *Biomedicine & Pharmacotherapy* **2024**, *175*, 116709.

(11) Marso, S. P.; Bain, S. C.; Consoli, A.; Eliaschewitz, F. G.; Jódar, E.; Leiter, L. A.; Lingvay, I.; Rosenstock, J.; Seufert, J.; Warren, M. L.; et al. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine* **2016**, *375*, 1834−1844.

(12) Husain, M.; Birkenfeld, A. L.; Donsmark, M.; Dungan, K.; Eliaschewitz, F. G.; Franco, D. R.; Jeppesen, O. K.; Lingvay, I.; Mosenzon, O.; Pedersen, S. D.; et al. Oral semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine* **2019**, *381*, 841−851.

(13) Wilding, J. P.; Batterham, R. L.; Calanna, S.; Davies, M.; Van Gaal, L. F.; Lingvay, I.; McGowan, B. M.; Rosenstock, J.; Tran, M. T.; Wadden, T. A.; et al. Once-weekly semaglutide in adults with overweight or obesity. *New England Journal of Medicine* **2021**, *384*, 989−1002.

(14) Vanhee, P.; van der Sloot, A. M.; Verschueren, E.; Serrano, L.; Rousseau, F.; Schymkowitz, J. Computational design of peptide ligands. *Trends Biotechnol.* **2011**, *29*, 231−239.

(15) Chang, L.; Mondal, A.; Perez, A. Towards rational computational peptide design. *Frontiers in Bioinformatics* **2022**, *2*, 1046493.

(16) Terada, T.; Inui, K.-i Recent advances in structural biology of peptide transporters. *Current topics in membranes* **2012**, *70*, 257−274.

(17) Miller, B. R.; Gulick, A. M. Structural biology of nonribosomal peptide synthetases. *Nonribosomal Peptide and Polyketide Biosynthesis: Methods and Protocols* **2016**, *1401*, 3−29.

(18) Stawikowski, M.; Fields, G. B. Introduction to peptide synthesis. *Current protocols in protein science* **2012**, *69*, 18.1.1−18.1.13.

(19) Erak, M.; Bellmann-Sickert, K.; Els-Heindl, S.; Beck-Sickinger, A. G. Peptide chemistry toolbox-Transforming natural peptides into peptide therapeutics. *Bioorganic & medicinal chemistry* **2018**, *26*, 2759−2765.

(20) Ferrazzano, L.; Catani, M.; Cavazzini, A.; Martelli, G.; Corbisiero, D.; Cantelmi, P.; Fantoni, T.; Mattellone, A.; De Luca, C.; Felletti, S.; et al. Sustainability in peptide chemistry: current synthesis and purification technologies and future challenges. *Green Chem.* **2022**, *24*, 975−1020.

(21) Wang, J.; Alekseenko, A.; Kozakov, D.; Miao, Y. Improved modeling of peptide-protein binding through global docking and accelerated molecular dynamics simulations. *Frontiers in molecular biosciences* **2019**, *6*, 112.

(22) Geng, H.; Chen, F.; Ye, J.; Jiang, F. Applications of molecular dynamics simulation in structure prediction of peptides and proteins. *Computational and structural biotechnology journal* **2019**, *17*, 1162−1170.

(23) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J. Struct. Biol.* **2007**, *157*, 593−605.

(24) Alam, N.; Goldstein, O.; Xia, B.; Porter, K. A.; Kozakov, D.; Schueler-Furman, O. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. *PLoS computation al biology* **2017**, *13*, e1005905.

(25) Zhang, Y.; Sanner, M. F. AutoDock CrankPep: combining folding and docking to predict protein−peptide complexes. *Bioinformatics* **2019**, *35*, 5121−5127.

(26) Chen, S.; Lin, T.; Basu, R.; Ritchey, J.; Wang, S.; Luo, Y.; Li, X.; Pei, D.; Kara, L. B.; Cheng, X. Design of target specific peptide inhibitors using generative deep learning and molecular dynamics simulations. *Nat. Commun.* **2024**, *15*, 1611.

(27) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic acids research* **2000**, *28*, 235−242.

(28) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(29) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(30) Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Venkatesh, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Anishchenko, I.; Humphreys, I. R.; et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **2024**, *384*, eadl2528.

(31) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.

(32) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein complex prediction with AlphaFold-Multimer. *biorxiv* **2021**, 2021.10.04.463034.
(33) Anishchenko, I.; Pellock, S. J.; Chidyausiku, T. M.; Ramelot, T. A.; Ovchinnikov, S.; Hao, J.; Bafna, K.; Norn, C.; Kang, A.; Bera, A. K.; et al. De novo protein design by deep network hallucination. *Nature* **2021**, *600*, 547−552.
(34) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I.; Courbet, A.; de Haas, R. J.; Bethel, N.; et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **2022**, *378*, 49−56.
(35) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089−1100.
(36) Ingraham, J. B.; Baranov, M.; Costello, Z.; Barber, K. W.; Wang, W.; Ismail, A.; Frappier, V.; Lord, D. M.; Ng-Thow-Hing, C.; Van Vlack, E. R.; et al. Illuminating protein space with a programmable generative model. *Nature* **2023**, *623*, 1070−1078.
(37) Fasman, G. D. *Prediction of protein structure and the principles of protein conformation*; Springer, 2012.
(38) Brooks, C.; Case, D. A. Simulations of peptide conformational dynamics and thermodynamics. *Chem. Rev.* **1993**, *93*, 2487−2502.
(39) Abagyan, R.; Argos, P. Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *Journal of molecular biology* **1992**, *225*, 519−532.
(40) Long, H. W.; Tycko, R. Biopolymer conformational distributions from solid-state NMR: α-helix and 310-helix contents of a helical peptide. *J. Am. Chem. Soc.* **1998**, *120*, 7039−7048.
(41) Lisanza, S. L.; Gershon, J. M.; Tipps, S. W. K.; Arnoldt, L.; Hendel, S.; Sims, J. N.; Li, X.; Baker, D. Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion. *bioRxiv* **2023**, 2023.05.08.539766.
(42) Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning* **2015**, 2256−2265.
(43) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **2020**, *33*, 6840−6851.
(44) Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv* **2020**, 2011.13456 DOI: 10.48550/arXiv.2011.13456.
(45) Kosugi, T.; Ohue, M. Solubility-aware protein binding peptide design using AlphaFold. *Biomedicines* **2022**, *10*, 1626.
(46) Yim, J.; Trippe, B. L.; De Bortoli, V.; Mathieu, E.; Doucet, A.; Barzilay, R.; Jaakkola, T. Se (3) diffusion model with application to protein backbone generation. *arXiv* **2023**, 2302.02277 DOI: 10.48550/arXiv.2302.02277.
(47) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind database: methodologies and updates. *Journal of medicinal chemistry* **2005**, *48*, 4111−4119.
(48) Martins, P. M.; Santos, L. H.; Mariano, D.; Queiroz, F. C.; Bastos, L. L. Propedia: a database for protein−peptide identification based on a hybrid clustering algorithm. *BMC bioinformatics* **2021**, *22*, 1−20.
(49) Wen, Z.; He, J.; Tao, H.; Huang, S.-Y. PepBDB: a comprehensive structural database of biological peptide-protein interactions. *Bioinformatics* **2019**, *35*, 175−177.
(50) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422.
(51) Zavala-Ruiz, Z.; Strug, I.; Walker, B. D.; Norris, P. J.; Stern, L. J. A hairpin turn in a class II MHC-bound peptide orients residues outside the binding groove for T cell recognition. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 13279−13284.
(52) Martin, S. E.; Tan, Z.-W.; Itkonen, H. M.; Duveau, D. Y.; Paulo, J. A.; Janetzko, J.; Boutz, P. L.; Törk, L.; Moss, F. A.; Thomas, C. J.; et al. Structure-based evolution of low nanomolar O-GlcNAc transferase inhibitors. *J. Am. Chem. Soc.* **2018**, *140*, 13542−13545.
(53) Pazgier, M.; Liu, M.; Zou, G.; Yuan, W.; Li, C.; Li, C.; Li, J.; Monbo, J.; Zella, D.; Tarasov, S. G.; et al. Structural basis for high-affinity peptide inhibition of p53 interactions with MDM2 and MDMX. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 4665−4670.
(54) Guerlavais, V.; Sawyer, T. K.; Carvajal, L.; Chang, Y. S.; Graves, B.; Ren, J.-G.; Sutton, D.; Olson, K. A.; Packman, K.; Darlak, K.; et al. Discovery of sulanemadlin (ALRN-6924), the first cell-permeating, stabilized α-helical peptide in clinical development. *J. Med. Chem.* **2023**, *66*, 9401−9417.
(55) Böttger, A.; Böttger, V.; Garcia-Echeverria, C.; Chène, P.; Hochkeppel, H.-K.; Sampson, W.; Ang, K.; Howard, S. F.; Picksley, S. M.; Lane, D. P. Molecular characterization of the hdm2-p53 interaction. *Journal of molecular biology* **1997**, *269*, 744−756.
(56) Klein, C.; Vassilev, L. Targeting the p53-MDM2 interaction to treat cancer. *British journal of cancer* **2004**, *91*, 1415−1419.
(57) Owen, D. R.; Allerton, C. M.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; et al. An oral SARS-CoV-2 Mpro inhibitor clinical candidate for the treatment of COVID-19. *Science* **2021**, *374*, 1586−1593.
(58) Mitchell, D.; Pobre, E. G.; Mulivor, A. W.; Grinberg, A. V.; Castonguay, R.; Monnell, T. E.; Solban, N.; Ucran, J. A.; Pearsall, R. S.; Underwood, K. W.; et al. ALK1-Fc inhibits multiple mediators of angiogenesis and suppresses tumor growth. *Molecular cancer therapeutics* **2010**, *9*, 379−388.
(59) Bendell, J. C.; Gordon, M. S.; Hurwitz, H. I.; Jones, S. F.; Mendelson, D. S.; Blobe, G. C.; Agarwal, N.; Condon, C. H.; Wilson, D.; Pearsall, A. E.; et al. Safety, pharmacokinetics, pharmacodynamics, and antitumor activity of dalantercept, an activin receptor-like kinase-1 ligand trap, in patients with advanced cancer. *Clin. Cancer Res.* **2014**, *20*, 480−489.
(60) Jang, D.-i.; Lee, A.-H.; Shin, H.-Y.; Song, H.-R.; Park, J.-H.; Kang, T.-B.; Lee, S.-R.; Yang, S.-H. The role of tumor necrosis factor alpha (TNF-α) in autoimmune disease and current TNF-α inhibitors in therapeutics. *International journal of molecular sciences* **2021**, *22*, 2719.
(61) Monaco, C.; Nanchahal, J.; Taylor, P.; Feldmann, M. Anti-TNF therapy: past, present and future. *International immunology* **2015**, *27*, 55−62.
(62) Rau, R. Adalimumab (a fully human anti-tumour necrosis factor α monoclonal antibody) in the treatment of active rheumatoid arthritis: the initial results of five trials. *Annals of the rheumatic diseases* **2002**, *61*, ii70−ii73.
(63) Xu, Y.; Wang, S.; Hu, Q.; Gao, S.; Ma, X.; Zhang, W.; Shen, Y.; Chen, F.; Lai, L.; Pei, J. CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic acids research* **2018**, *46*, W374−W379.
(64) Wang, S.; Xie, J.; Pei, J.; Lai, L. CavityPlus 2022 update: an integrated platform for comprehensive protein cavity detection and property analyses with user-friendly tools and cavity databases. *J. Mol. Biol.* **2023**, *435*, 168141.
(65) Miura, T.; Malla, T. R.; Owen, C. D.; Tumber, A.; Brewitz, L.; McDonough, M. A.; Salah, E.; Terasaka, N.; Katoh, T.; Lukacik, P.; et al. In vitro selection of macrocyclic peptide inhibitors containing cyclic γ2, 4-amino acids targeting the SARS-CoV-2 main protease. *Nat. Chem.* **2023**, *15*, 998−1005.
(66) Salmon, R. M.; Guo, J.; Wood, J. H.; Tong, Z.; Beech, J. S.; Lawera, A.; Yu, M.; Grainger, D. J.; Reckless, J.; Morrell, N. W.; et al. Molecular basis of ALK1-mediated signalling by BMP9/BMP10 and their prodomain-bound forms. *Nat. Commun.* **2020**, *11*, 1621.
(67) Das, R.; Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **2008**, *77*, 363−382.
(68) McMillan, D.; Martinez-Fleites, C.; Porter, J.; Fox, D., 3rd; Davis, R.; Mori, P.; Ceska, T.; Carrington, B.; Lawson, A.; Bourne, T.; et al. Structural insights into the disruption of tnf-tnfr1 signalling by

small molecules stabilising a distorted tnf. *Nat. Commun.* **2021**, *12*, 582.
(69) Al Musaimi, O.; Lombardi, L.; Williams, D. R.; Albericio, F. Strategies for improving peptide stability and delivery. *Pharmaceuticals* **2022**, *15*, 1283.
(70) Valdés-Tresanco, M. S.; Valdés-Tresanco, M. E.; Valiente, P. A.; Moreno, E. gmx_MMPBSA: a new tool to perform end-state free energy calculations with gromacs. *J. Chem. Theory Comput.* **2021**, *17*, 6281−6291.
(71) Miller, B. R., III; McGee, T. D., Jr; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA. py: an efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* **2012**, *8*, 3314−3321.
(72) Zhang, J.; Durham, J.; Cong, Q. Revolutionizing protein-protein interaction prediction with deep learning. *Curr. Opin. Struct. Biol.* **2024**, *85*, 102775.
(73) Notin, P.; Rollins, N.; Gal, Y.; Sander, C.; Marks, D. Machine learning for functional protein design. *Nat. Biotechnol.* **2024**, *42*, 216−228.
(74) Gupta, S.; Azadvari, N.; Hosseinzadeh, P. Design of protein segments and peptides for binding to protein targets. *BioDesign Research* **2022**, *2022*, 9783197.
(75) Wang, Y.-C.; Bai, S.-C.; Ye, W.-L.; Jiang, J.; Li, G. Recent progress in site-selective modification of peptides and proteins using macrocycles. *Bioconjugate Chem.* **2024**, *35*, 277.
(76) Scantlebury, J.; Brown, N.; Von Delft, F.; Deane, C. M. Data set augmentation allows deep learning-based virtual screening to better generalize to unseen target classes and highlight important binding interactions. *J. Chem. Inf. Model.* **2020**, *60*, 3722−3730.
(77) Gao, B.; Jia, Y.; Mo, Y.; Ni, Y.; Ma, W.; Ma, Z.; Lan, Y. Self-supervised pocket pretraining via protein fragment-surroundings alignment. *arXiv* **2023**, 2310.07229.
(78) Huang, H.; Sun, L.; Du, B.; Lv, W. Learning joint 2d & 3d diffusion models for complete molecule generation. *arXiv* **2023**, 2305.12347.
(79) Le, T.; Cremer, J.; Noé, F.; Clevert, D.-A.; Schütt, K. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation. *arXiv* **2023**, 2309.17296.
(80) Anand, N.; Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv* **2022**, 2205.15019.
(81) Campbell, A.; Yim, J.; Barzilay, R.; Rainforth, T.; Jaakkola, T. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv* **2024**, 2402.04997.
(82) Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow matching for generative modeling. *arXiv* **2022**, 2210.02747.
(83) Chen, R. T.; Lipman, Y. Riemannian flow matching on general geometries. *arXiv* **2023**, 2302.03660.
(84) Mannhold, R., Kubinyi, H., Folkers, G. Pharmacophores and pharmacophore searches. Wiley Online Library, 2006.
(85) Weikl, T. R.; Paul, F. Conformational selection in protein binding and function. *Protein Sci.* **2014**, *23*, 1508−1518.
(86) Zarutskie, J. A.; Sato, A. K.; Rushe, M. M.; Chan, I. C.; Lomakin, A.; Benedek, G. B.; Stern, L. J. A conformational change in the human major histocompatibility complex protein HLA-DR1 induced by peptide binding. *Biochemistry* **1999**, *38*, 5878−5887.
(87) Springer, S.; Döring, K.; Skipper, J. C.; Townsend, A. R.; Cerundolo, V. Fast association rates suggest a conformational change in the MHC class I molecule H-2Db upon peptide binding. *Biochemistry* **1998**, *37*, 3001−3012.
(88) Armstrong, K. M.; Piepenbrink, K. H.; Baker, B. M. Conformational changes and flexibility in t-cell receptor recognition of peptide−mhc complexes. *Biochem. J.* **2008**, *415*, 183−196.
(89) Malde, A. K.; Hill, T. A.; Iyer, A.; Fairlie, D. P. Crystal structures of protein-bound cyclic peptides. *Chem. Rev.* **2019**, *119*, 9861−9914.
(90) Nielsen, D. S.; Shepherd, N. E.; Xu, W.; Lucke, A. J.; Stoermer, M. J.; Fairlie, D. P. Orally absorbed cyclic peptides. *Chem. Rev.* **2017**, *117*, 8094−8128.
(91) Gavenonis, J.; Sheneman, B. A.; Siegert, T. R.; Eshelman, M. R.; Kritzer, J. A. Comprehensive analysis of loops at protein-protein interfaces for macrocycle design. *Nat. Chem. Biol.* **2014**, *10*, 716−722.
(92) Demmer, O.; Frank, A. O.; Kessler, H. Design of cyclic peptides. *Peptide and protein design for biopharmaceutical applications* **2009**, 133−176.
(93) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **2017**, *13*, e1005659.
(94) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150−3152.
(95) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658−1659.
(96) Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **2010**, *26*, 680−682.
(97) Köhler, J.; Klein, L.; Noé, F. Equivariant flows: exact likelihood generative learning for symmetric densities. *International conference on machine learning* **2020**, 5361−5370.
(98) Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; Tang, J. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv* **2022**, 2203.02923.
(99) De Bortoli, V.; Mathieu, E.; Hutchinson, M.; Thornton, J.; Teh, Y. W.; Doucet, A. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems* **2022**, *35*, 2406−2422.
(100) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems. *Advances in neural information processing systems* **2017**, 30.
(101) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* **2022**, *2022*, 2022.07.20.500902.
(102) Chen, T.; Zhang, R.; Hinton, G. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv* **2022**, 2208.04202.
(103) Ba, J. L.; Kiros, J. R.; Hinton, G. E. Layer normalization. *arXiv* **2016**, 1607.06450.
(104) Conway, P.; Tyka, M. D.; DiMaio, F.; Konerding, D. E.; Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein science* **2014**, *23*, 47−55.
(105) Dauparas, J.; Lee, G. R.; Pecoraro, R.; An, L.; Anishchenko, I.; Glasscock, C.; Baker, D. Atomic context-conditioned protein sequence design using LigandMPNN. *Biorxiv* **2023**, 2023.12.22.573103.
(106) Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, 1711.05101.
(107) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* **2019**, *32*, 1.
(108) Bennett, N. R.; Coventry, B.; Goreshnik, I.; Huang, B.; Allen, A.; Vafeados, D.; Peng, Y. P.; Dauparas, J.; Baek, M.; Stewart, L.; et al. Improving de novo protein binder design with deep learning. *Nat. Commun.* **2023**, *14*, 2625.

https://doi.org/10.1021/acs.jcim.4c00975
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX